INTRODUCTION TO MACHINE LEARNING COMPSCI 4ML3

LECTURE 11

HASSAN ASHTIANI

MAXIMUM LIKELIHOOD ESTIMATE

- MAXIMIZES THE PROBABILITY OF THE OBSERVATIONS GIVEN THE PARAMETERS
  - $\alpha^{ML} = \underset{\alpha}{argmax}\, P\,(X|\alpha)$
  - $\alpha^{ML} = \underset{\alpha}{argmin} -\left(\sum_i \text{LOG}\, P\,(x^i|\alpha)\right)$
- FOR BIAS OF THE COIN
  - $\alpha^{ML} = \frac{n_0}{n_0+n_1}$

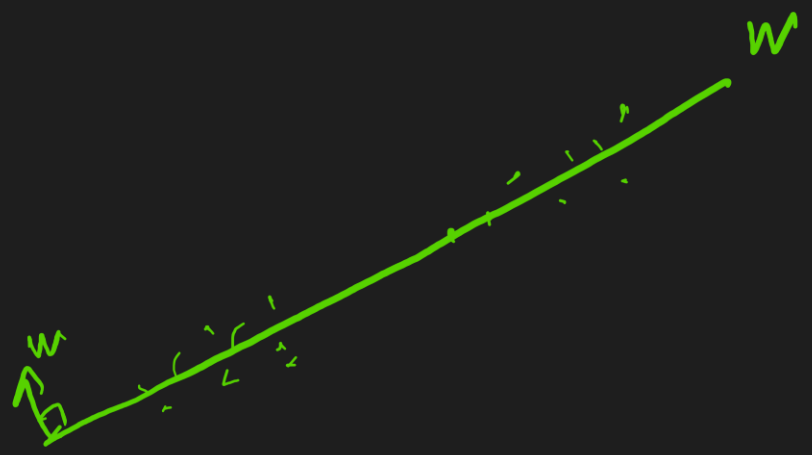
MAXIMUM A POSTERIORI ESTIMATE

- MAXIMIZES THE PROBABILITY OF THE PARAMETERS GIVEN THE OBSERVATIONS
  - $\boldsymbol{\alpha^{MAP} = \underset{\alpha}{argmax}\, P\,(\alpha|X)}$ $= \arg\max P(X|\alpha) \cdot P(\alpha)$
- $\alpha^{MAP} = \underset{\alpha}{argmin}\left(-\text{LOG}(P(\alpha)) - \sum_{i=1}^{n} \text{LOG}\, \boldsymbol{P}(x^i|\alpha)\right)$

PRIOR VS POSTERIOR DISTRIBUTIONS

- $P(\alpha)$ CAPTURES THE **PRIOR** DISTRIBUTION
- $P(\alpha|X)$ CAPTURES THE **POSTERIOR** DISTRIBUTION
- IN OTHER WORDS,
  - WE START BY A PRIOR BELIEF ABOUT VALUE OF α
  - OUR BELIEF IS UPDATED AFTER SEEING SOME REAL DATA
  - THIS IS A BAYESIAN APPROACH

- $P(y|x, W) = \left(\frac{1}{\gamma}\right) e^{-\frac{\left(x^T W - y\right)^2}{2\sigma^2}}$
- $P(x|W) = P(x)$



- INPUT: $((x^1, y^1), \ldots, (x^n, y^n))$ I.I.D. SAMPLE
  - EACH (x^i, y^i) IS DRAWN ACCORDING TO P(X,Y)
- MAXIMUM LIKELIHOOD ESTIMATE FOR W?

ML ESTIMATE

$$\arg\max_w p(z|\alpha) = \arg\max_w \sum \log p(x^i, y^i | w)$$

$$= \arg\max_w \sum \log \left[\frac{p(x^i, y^i, w)}{p(w)} \right]$$

$$= \arg\max_w \sum \log \left[p(y^i | x^i, w)\, p(x^i | w)\, p(w) / p(w) \right]$$

$$= \arg\max_w \sum \left[\left(\ln \frac{1}{\sigma} \right) - \frac{(x^{iT} w - y^i)^2}{\sigma^2} + \ln p(x^i) \right]$$

$$= \arg\min_w \sum_i (x^{iT} w - y^i)^2$$

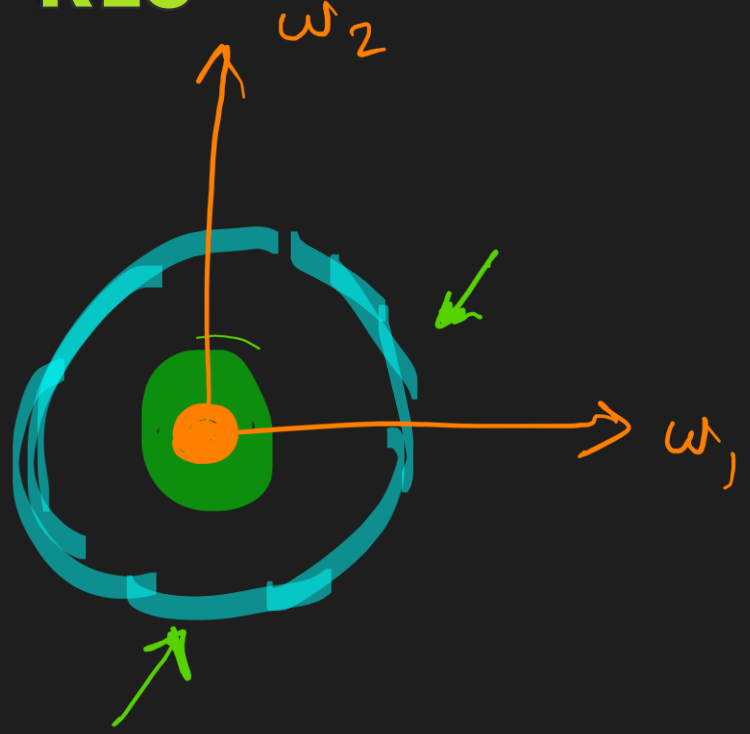
A PROBABILISTIC MODEL FOR REGRESSION

- $P(y|x, W) = \left(\frac{1}{\gamma}\right) e^{-\frac{\left(x^T W - y\right)^2}{2\sigma^2}}$
- $P(x|W) = P(x)$

- MAXIMUM LIKELIHOOD SOLUTION FOR THIS PROBABILISTIC MODEL IS THE SAME AS LS SOLUTION.

MAXIMUM A POSTERIORI AND RLS

- $P(y|x, W) = \left(\frac{1}{\gamma}\right) e^{-\frac{\left(x^T W - y\right)^2}{2\sigma_1^2}}$
- $P(x|W) = P(x)$
- $P(W) = \frac{1}{\beta} e^{-\lambda \|W\|_2^2}$



what if we have

$$p(w) = \frac{1}{\beta} e^{\frac{-\lambda \|w\|_2^2}{\sigma_2^2}} \; ?$$

MAP ESTIMATE

$$\arg\max_w p(x|\alpha)p(\alpha) = \arg\max_w \left[\log p(\alpha) + \sum_i \log(x^i, y^i | w) \right]$$

$$= \underline{\qquad\qquad}$$

$$= \arg\max_w \left[\ln \frac{1}{p} - \lambda \|w\|_2^2 - \frac{(x^{iT} w - y^i)^2}{\sigma^2} \right]$$

$$= \arg\min_w \left[\sum_i (x^i w - y^i)^2 + \lambda \sigma^2 \|w\|_2^2 \right]$$

e.g., if $\sigma = 1$ then we get the standard RLS.

EXPECTED ERROR MINIMIZATION

- SQUARED LOSS
  - $l(\hat{y}, y) = (y - \hat{y})^2$
- EMPIRICAL/TRAINING LOSS FOR $Z = \{(x^i, y^i)\}_{i=1}^n$

$$\frac{1}{n} \sum (y^i - \hat{y}^i)^2$$

- EXPECTED LOSS

$$L = E_{(x,y) \sim P(x,y)} l(\hat{y}(x), y)$$

REWRITING THE EXPECTATION

$$L = \mathop{E}_{(x,y)\sim p(x,y)} f(x,y) = \mathop{E}_{x\sim p(x)} \mathop{E}_{y\sim p(y|x)} f(x,y)$$

(Assuming the expectation exist)

Chain Rule

EXPECTED ERROR MINIMIZATION

- ASSUME WE KNOW $P(x, y)$
  - $\boldsymbol{y^*(x) = argmin_{\hat{y}} E_{x,y}(y - \hat{y}(x))^2 = ?}$

$$\underset{\hat{y}(\cdot)}{\operatorname{argmin}} \; \underset{x \sim P(x)}{E} \; \underset{y \sim P(y|x)}{E} \; (y - \hat{y}(x))^2$$

you can choose $\hat{y}(x)$ based on x

$$\rightarrow \underset{\hat{y}(x) \in \mathbb{R}}{\operatorname{argmin}} \; \underset{y \sim P(y|x)}{E} \; (y - \hat{y}(x))^2 \; ?$$

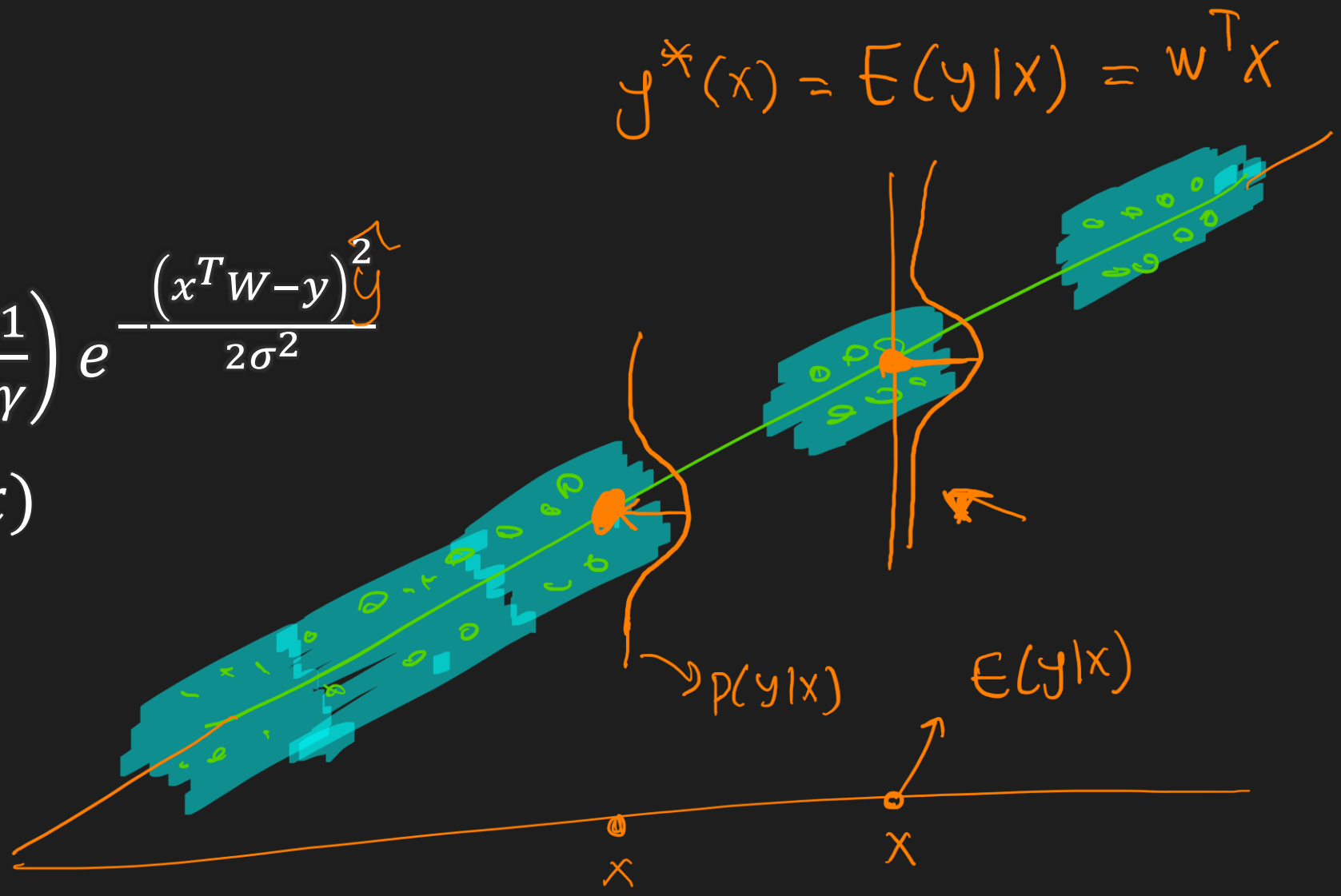
$$\frac{\partial}{\partial \hat{y}(x)} \underset{y \sim p(y|x)}{\mathbb{E}} \left(\hat{y}(x) - y \right)^2$$ where x is given.

$$= 2 \underset{y \sim p(y|x)}{\mathbb{E}} \left(\hat{y}(x) - y \right) = 0$$

$$\longrightarrow \hat{y}(x) = \underset{y \sim p(y|x)}{\mathbb{E}} y = \mathbb{E}[y|x]$$

EXAMPLE

- $P(y|x, W) = \left(\frac{1}{\gamma}\right) e^{-\frac{(x^T W - y)^2}{2\sigma^2}}$
- $P(x|W) = P(x)$

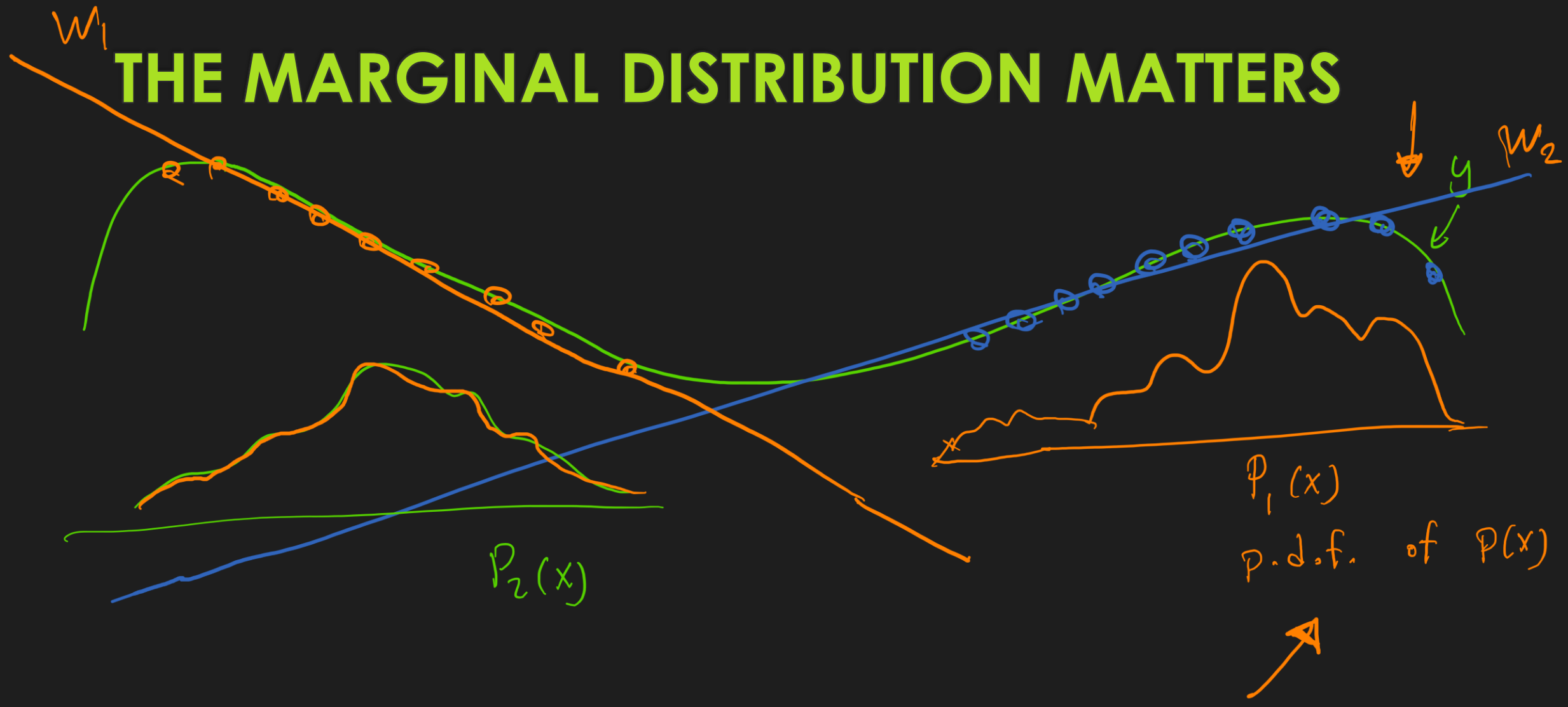


THE MARGINAL DISTRIBUTION MATTERS

- In practice, we don't know $P(X, Y)$
- Also, it is hard to estimate $E(Y|X)$ for every X
  - But we can aim to estimate $E(Y|X)$ "where it matters"

- Even for the same $P(Y|X)$, W_1 or W_2 may compare differently depending on $P(X)$
  - Example?

THE MARGINAL DISTRIBUTION MATTERS



~~Although~~ Despite $p(y|x)$ was the same.